Biom. Praxim. (1984), 24, 49-65

OVERLAPPING RECURRENT GROUPS : AN EXTENSION OF FAGER'S CONCEPT AND ALGORITHM

H.M. ANDRE

Université Catholique de Louvain
Laboratoire d'Ecologie et de Biogéographie
Place Croix du Sud 5
B-1348 Louvain-la-Neuve, Belgium

SUMMARY

The biological meaning of recurrent groups sensu *FAGER (1957) is reviewed and briefly discussed. A formal definition is subsequently proposed. As an extension of FAGER'S procedure, a new algorithm is presented which leads to the formation of overlapping recurrent groups. The new procedure, implemented in a FORTRAN program called OVRGROUP, is applied to dummy data. This highlights major differences between the old and new procedures.*

1. INTRODUCTION

If two species are affected by the same environmental factors, if they search for the same locations (oviposition sites, feeding sites, etc), or if they have some interactions, either favorable or unfavorable, on each other, their spatial distribution will not be independent. If it turns out that these two species are found together frequently, they will be said to be *co-occurring*. In other words, such species are frequent members of each other's environment and grouping them together seems to be cogent. This outlines the basic approach chosen by FAGER (1957) to define his ecological "recurrent groups".

Such an approach is not new and the notion of recurrent groups is close to that of "association" advanced by the Zurich-Montpellier School in plant sociology. In particular, it agrees with the concept of association proposed by BROCKMANN-JEROSCH (1907) and GRADMANN (1909) (both quoted by WHITTAKER (1962) and SHIMWELL (1971)) and developed by

BRAUN-BLANQUET as soon as in 1913. In that context, these plant sociologists propounded the term of character species to designate species occurring with a great constancy in a particular vegetation type.

On the other hand, the concept of recurrent groups somewhat accords with that of *synusia* introduced by GAMS (1918) and defined as a unit composed of species of similar life forms *with similar ecological requirements* in a broad sense. It also approximates the concept of "ecological group" as defined by DUVIGNEAUD (1946) but it differs somewhat from ELLENBERG'S (1950) ecological groups. The former author emphasizes the ecosociological affinities while the latter stresses the similarity of responses to the environment and disregards sociological affinity.

Similar views to FAGER'S opinion have also been expressed by numerous zoologists, for instance GISIN (1946), MARGALEF (1968), KÜHNELT (1969)...

This (too) brief review of the literature reveals that several concepts are misty and/or overlapping; as a result, the situation is rather confusing. The objective of this paper is to start from a biological definition of recurrent group, to draw up a more formal definition and then to build up an unsophisticated algorithm, consistent with that definition and likely to determine such groups. The properties of the resulting algorithm, OVRGROUP, will be compared to those of a previous and similar algorithm, namely REGROUP, the procedure advanced by FAGER (1957), and subsidiarily, to those of KDEND, a program proposed by COLE and WISHART (1970). As an example, the new algorithm is applied to dummy data; application to field data is displayed in a subsequent paper (ANDRE, 1985).

2. *THE RECURRENT GROUP : CONCEPT AND DEFINITION*

The expression "recurrent group" was not formally defined by FAGER (1957). He just provided a set of four requirements which a group must satisfy. In fact, those requirements are rules governing the determination of groups and describe the classification algorithm. FAGER and McGOWAN (1963) do not give a formal definition either but they stress that the study of their data has shown that *certain species frequently occur together* and that these *species groups characterize* particular habitats. They conclude that the groups should be composed of species that have *similar reactions to properties of the environment* (food, microclimate...). The last point is also emphasized by VENRICK (1971).

A fundamental definition comes out from the above literature: a recurrent group consists of species occurring frequently together. As an extension of this definition, the concept suggests two properties :

- a recurrent group characterizes a particular habitat,
- a recurrent group consists of species showing similar reactions to characteristics of the environment or having similar ecological requirements.

To draw up an efficient algorithm, we now need a formal definition. Let be P the set of species. Thus, a recurrent group, G, is a given relation (subset), $R(P^2)$, defined on the Cartesian product, P^2, together with a mapping $g \rightarrow a(g)$ assigning an affinity index, $a(g)$, greater or equal to a treshold value T, to every element g of R; *i.e.* (using KLIR'S (1969) notation) :

either $G = R(P^2) \subseteq P^2$ (1),

$$G = \{R(P^2),\ g \rightarrow a(g)\ :\ \forall g \in R,\ a(g) \geqslant T\} \quad (2).$$

Such a definition means and implies that

- each species belonging to the group G has a positive affinity with every other species of the group;
- the affinity index, $a(g)$, between any two species of the group G has to be greater or equal to a treshold value T.

In this context, affinity must be taken in a mathematical sense. Affinity indices, also called indices or coefficient of association, measure the degre of coexistence between two species over different habitats or localities. High values of such indices do not imply a biological affinity or any kind of relationships as parasitism or episitism but merely indicate that two species are very often co-occurring.

On the other hand, the above definition allows several groups satisfying the definition to be extracted from a data matrix. Contrary to the algorithm developed by FAGER (1957), it does not preclude that a species belongs to several groups.

Now, to make efficient the above definition, it is necessary to

- define an affinity index, a(g) ;
- determine the treshold value, T ;
- specify how to extract the successive groups from the data; *i.e.* how to classify the species.

3. *AN AFFINITY INDEX,* α.

According to the above definition, the affinity index must rest on presence and absence data, *i.e.* on binary data. In this context, it is usual to distinguish two questions. On the one hand, is the affinity between two species significant? In other words, are the two species spatially independent of each other? On the other hand, how strong is that association if it exists?

Testing the affinity between two species seems to be simple and numerous ecology textbooks still propose to do it by applying a chi-square test to a 2 x 2 contingency table(*). As early as in 1957, Fager criticized this method. Table I comprises three series of examples which clearly illustrate the irrelevancy of the chi-square test. Within each series, the total number of samples is constant as is the number of empty samples (joint or mutual absences, cell d in the contingency table). Any ecologist looking at the first series would assert that the affinity is increasing from left to right; however, the computed value of χ^2 is minimal in example B. In the second series, the pattern is the same: the affinity increases from left to right; in this case, the chi-square test supplies cogent values; in particular, the affinity is not significant in

---

(*) The notation used in this paper will follow that of a 2 x 2 contingency table where

a = number of releves in which species 1 and 2 both occur;
b = number of releves in which species 2 occurs alone;
c = number of releves in which species 1 occurs alone;
d = number of releves in which both species are absent.

example D. The third series is similar to the second one but there are no empty samples (mutual absence). Although the other values are the same as in the second series, the chi-square values are decreasing from left to right and, in case I, it is not significant at a 0.01 level.

Lastly, in any contingency table, cells "a" and "d", *i.e.* respectively "mutual presence" and "mutual absence" can be inverted without changing the result ot the chi-square test ! These surprising results are however easy to understand: the chi-square test supplies a measure of the discrepancy existing between observed and expected frequencies related to the total number of samples (a + b + c + d). Example B clearly shows that relative proportions are equal within each column and within each row and thus observed frequencies are equal to expected ones.

It results that the chi-square test and derived index as that proposed by KRYLOV (1968) cannot be used to test the affinity between two species.

FAGER and McGOWAN (1963) have thus presented a new approach. First, they have proposed a new index. Using the notation for a 2 x 2 contingency table, it is :

$$\alpha = \frac{a}{\sqrt{(a+b)(a+c)}} - \frac{1}{2\sqrt{\max\{(a+b),(a+c)\}}} \qquad (3)$$

This index has several desirable properties :

- The probabilities of occurrences of the two species are related to the sum of their occurrences rather than to the total number of samples; this means that joint absences are disregarded.
- The index defined above is close to OCHIAI (1957) index, the "measure of ecological coexistence" preferred by JANSON and VEGELIUS (1981); both correspond to the geometric mean of the proportion of joint occurrences but the index above is corrected for sample size.
- The use of α is encompassed in a more general procedure (see appendix) which allows a treshold value to be determined as explained in the next section. However ,it must be admitted

that the procedure described hereafter could be used with any other association index (for a critical review, see JANSON and VEGELIUS (1981) or HUBÁLEK (1982)).

TABLE I : Three series of three 2 x 2 contingency tables. Corresponding values of the χ^2-test and α are displayed below each table. The strength of the association increases from left to right and is similar in the 2nd and 3rd series.

A.

380	160	540
160	20	180
540	180	720

$\chi^2 = 24.69$; $\alpha = .68$

B.

500	100	600
100	20	120
600	120	720

$\chi^2 = 0$; $\alpha = .81$

C.

620	40	660
40	20	60
660	60	720

$\chi^2 = 53.55$; $\alpha = .92$

D.

134	134	268
134	100	234
268	234	502

$\chi^2 = 2.65$; $\alpha = .47$

E.

202	100	302
100	100	200
302	200	502

$\chi^2 = 14.32$; $\alpha = .64$

F.

312	45	357
45	100	142
357	145	502

$\chi^2 = 159.46$; $\alpha = .85$

G.

134	134	268
134	0	134
268	134	402

$\chi^2 = 100.50$; $\alpha = .47$

H.

202	100	302
100	0	100
302	100	402

$\chi^2 = 44.08$; $\alpha = .64$

I.

312	45	357
45	0	45
357	45	402

$\chi^2 = 6.39$; $\alpha = .85$

4. *THE TRESHOLD VALUE, T.*

As well noticed by KRYLOV (1968), the calculation of indices measuring the strength of an association is only meaningful if the existence of the association has been established. At first sight, a problem arises because the above index, α, does not give any information on a significant treshold value.

However, a program designed by FAGER and distributed by the Scripps Institution of Oceanography provides this kind of information. The AFFIN program calculates affinity indices for species assigned *at random* to set a samples. The number and frequencies of species and the number of samples exactly coincide to those observed in the real data to be analyzed. In other words, the program simulates a random distribution of the species collected in the field and then forms a cumulative frequency distribution of the index values. From this, it easy to determine what level the index should be set at in order to have a mean chance of 95 or 99 %, over the entire suite of species, of getting that high or higher an index if the species were randomly distributed among samples. However, as outlined by VENRICK (*in litt.*), the value of T computed from program AFFIN does not determine the significance of any recurrent group formed at that treshold value. Practically speaking, it does not mean that *a* recurrent group formed with that treshold value will have a probability of occurring by chance alone (*i. e.* if the species were in fact independant) of 0.01 or 0.05. Nevertheless, AFFIN does provide an objective way of chosing comparable levels of α for different sets of samples. Hereafter, two species characterized by an affinity index value greater than the treshold value T will be said "linked" or "interrelated". In other words, there is a "link" or a "bond" between them.

5. *THE PREVIOUS CLASSIFICATION ALGORITHMS*

Its results from the formal definition proposed above that a complete linkage algorithm has to be chosen. FAGER (1957) has presented a classification method he implemented in a FORTRAN program called REGROUP. This program uses, α, the index of FAGER and McGOWAN (1963) defined in equation (3) and rests essentially on two rules :

1. A group includes the greatest possible number of interrelated species.
2. If several groups with the same number of species are possible, that is selected which has the greatest amount of affinity values between its species.

The classification procedure is summarized in three points :

Step 1 : The largest possible group of interrelated spécies is isolated; if there are several possible groups, rule 2 applies.

Step 2 : The "bonds" or "links" of the species belonging to that group are deleted, *i.e.* their affinity values are set to zero.

Step 3 : Back to step 1.

This algorithm prompts some comments.
Rule 1, and thus step 1 of the procedure, is not deduced from the definition of recurrent group. However from a practical point of view it seems difficult to disregard it. In addition, splitting a group comprising a maximum number of species into several subgroups is not necessarily meaningfull or desirable from an ecological point of view.

The procedure is an iterative process (see step 3) which is applied on the *remaining* species, *i.e.* on species that do not yet belong to a group already defined in a previous iteration (see step 2). The number, n_i, of remaining species (i.e. still to be classified) after any iteration, i, is thus

$$n_i = N - \Sigma_{j=1}^{i} S_j \qquad (4)$$

where N is the total number of species and S_j the number of species in the jth group.
This procedure implies that the recurrent groups are mutually exclusive and that a species cannot belong to two or more

groups. Once again, this is not deduced from the above mentioned definition.

The second remark is more serious than it could be expected in the sense that FAGER's algorithm leads to a systematic bias. Indeed, the first group determined by such an algorithm comprises nearly always a real group plus some "ubiquitous" species which are no more associated with species of the first group than with others belonging to subsequent groups.

This is clearly illustrated when REGROUP is applied to dummy data as those presented in figure 1. Figure 1A and 1D show two sets of data. Each of them comprises 25 samples. In both examples, two groups of species are easily distinguished. In the first case, species e is a transition species, in the second case, species e is ubiquitous.
Corresponding graphs or species-association plexus indicating affinities greater than .5(*) are shown in figures 1B and 1E. In both cases, FAGER's algorithm puts species e into the first group (with species b, c and d) and, next, groups species f, g, h and i. The bias introduced by the procedure is obvious.

A nearly identical procedure has been advanced by KRYLOV (1968) who introduced a new rule : if several groups with the same number of species are possible, that is selected which contains the species having associations with the lesser number of species outside the group. In other words, KRYLOV' criterion selects groups characterized by a minimum number of external bounds. As noticed by KRYLOV, preference is thus given to species more highly "specialized" in respect to their group. This algorithm, also used by LEGENDRE and BEAUVAIS (1978), leads often to the same bias as FAGER's procedure.

On the other hand, KRYLOV (1968) pointed out it is meaningless to select groups of two species. Indeed, such pairs are not selected because of a particularly high value of α between both species but merely because it is impossible to find a larger group. Therefore the procedure he proposed stops

---

(*) This arbitrary level coincides with that selected by FAGER and McGOWAN (1963).

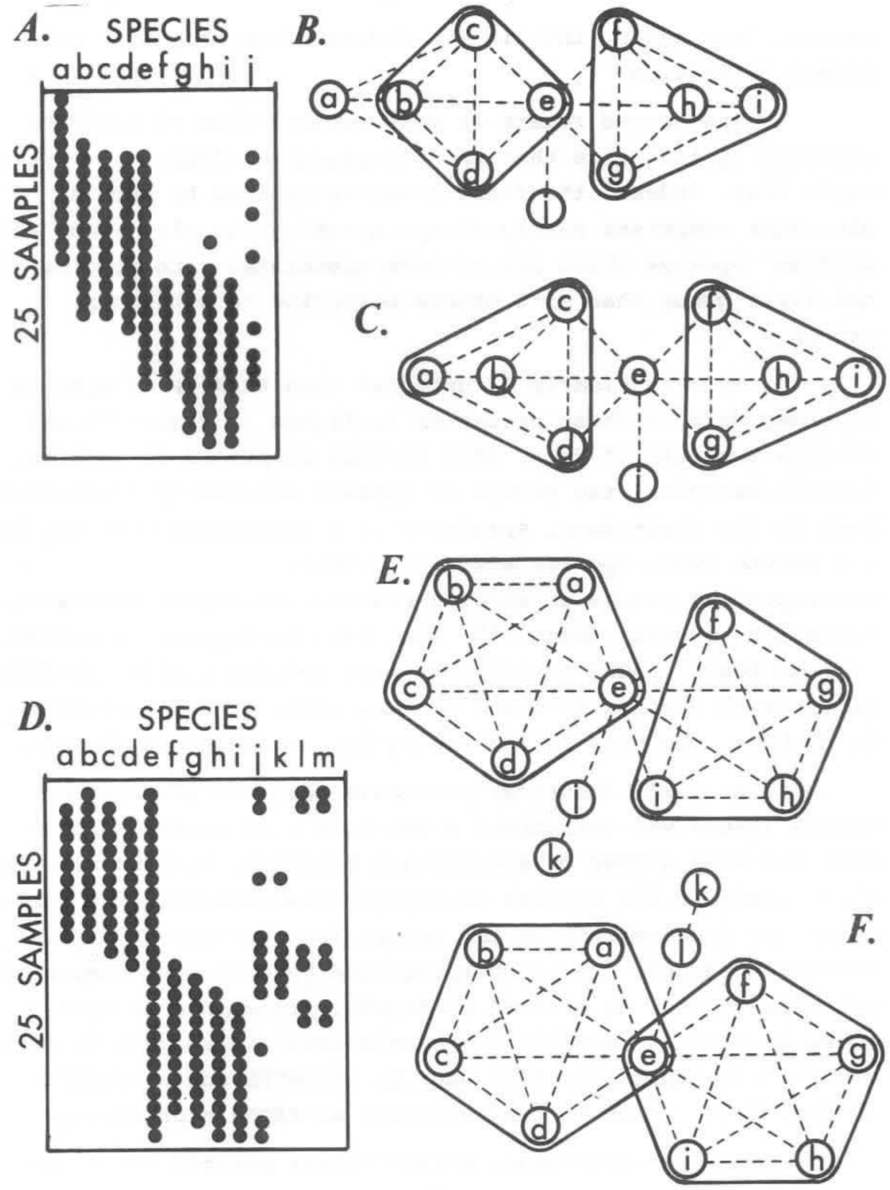

Figure 1 - Two presence/absence matrices of 25 relevés (A, D) with the corresponding graphs (B, C; E, F). Recurrent groups formed with REGROUP (B, E) and OVRGROUP (C, F) are outlined.

when it is no longer possible to find a new group of at least three species. In contrast, FAGER's procedure terminates only when no group of two species can be formed.

6. A NEW ALGORITHM : DESCRIPTION OF OVRGROUP

To overcome the bias illustrated in the previous section, a new algorithm has been drawn up and implemented in a program called OVRGROUP and which also uses the index of FAGER and McGOWAN (1963). Before describing it, it must be emphasized that searching for recurrent groups among a biocenosis is tantamount to isolating complete subgraphs in the graph or species-association plexus representing the whole biocenosis (as illustrated in figures 1B and C). Putting the problem in the frame of the graph theory seems thus specially cogent as stressed by DALE (1977). Main rules implemented in OVRGROUP are briefly described hereafter :

Rule 1

A group includes the greatest possible number of interrelated species. Following the graph theory terminology exposed by DALE (1977), this is tantamount to searching for cliques, *e.i.* maximal complete subgraphs. Note that this rule was also implemented in previous algorithm.

Rule 2

If several groups with the same number of species are possible, OVRGROUP offers two possibilities. Either the user chooses FAGER's criterion and selects groups with the greatest amount of affinity values between its species, or he chooses KRYLOV's criterion and selects group with minimum number of external bonds (if there are still several possible groups, then FAGER's criterion applies).

Rule 3

Equation (4) above determining the number of species still to be classified after each iteration does not hold. A fundamental distinction is made between introverted species in a group and extroverted ones. The former is defined as species having more links with species of its own group than with others; *i.e.* having more internal links than external. In other words, the degree(*) of a point representing an introverted species must be less than twice the group size.

If a group contains several introverted species, they are handled the same way as in FAGER's procedure and belong to one and only one group. They are *indicator species*. If a group contains only one introverted species, it remains temporarily in the set of species to be classified until it is put in one and only one other group in a subsequent iteration; if that species is not put in a second group at the end of the procedure, it is also said to be an indicator species.

In contrast, extroverted species have as many or more external links than internal and they remain permanently in the set of species to classify. They are thus likely to belong to several groups; if they do not belong to several groups, they are also considered to be indicator species at the end of the procedure.

Rule 3 prompts some comments, especially about the way a single introverted species is handled. Such a species is a special case intermediate between the other introverted species which are immediately recognized as indicator species, and the extroverted species whose status is specified only at the end of the procedure. A recurrent group comprising only extroverted species but one contains euryecious or ubiquitous species whose habitat is not necessarily well defined. Consequently, the group habitat itself is more or less undefined and it is felt that a single introverted species cannot characterize alone such a group unless it is recognized as an indicator species at the end of the procedure. This is why it is given a single introverted species one more chance to belong to another group. The effects of this special handling do not appear with the dummy data hereafter but will be discussed in the subsequent paper dealing with field data (ANDRE, 1985, appendix III).

(*) In graph theory, the degree of a point p is the number of points to which p is connected.

Rule 4

A group cannot be a subset of the set of extroverted species. Due to the preceding rules, the danger is that the computer enters an endless loop by selecting always the same group of extroverted species. Rule 4 prevents such a possibility.

Rule 5

Groups of two species are not selected except if the sum of the degrees of points representing them is equal to two or three. Such situations respectively corresponds to

a) isolated pairs of species, *i.e.* disconnected from the rest of the graph, and to

b) pairs forming a lateral branch linked to the rest of the graph through a bridge(*).

Rule 6

OVRGROUP has a special feature which, after a group is determined, searches for species which are linked only to species forming that group. This corresponds to a single linkage procedure and such species are said associate with the group.

7. *APPLICATION TO DUMMY DATA*

The new algorithm was applied to the data represented in figure 1, which also shows the resulting groups. Recurrent groups obtained with OVRGROUP (fig. 1C, F) are overlapping and appear to be more consistent with the expected classification than those obtained with REGROUP (fig. 1B, E). In particular, the special status of species e is well displayed. In figure 1C, species e appears to be a transition species between both groups while, in figure 1F, it turns out to be an ubiquitous species belonging to both groups.

(*) In graph theory, an edge is a bridge if its removal disconnects the graph.

Those examples reveal the fundamental difference between the algorithms; they are however too simple to detail how OVRGROUP works in real situations. Therefore, OVRGROUP has been applied to field data together with REGROUP. Comparative results are detailed in a subsequent paper (ANDRE, 1985).

8. CONCLUSIONS

In a general context, OVRGROUP could be interpreted as an intermediate tool between REGROUP which, in a graph, searches for exclusive maximal complete subgraphs and KDEND (COLE and WISHART, 1970; WISHART, 1978) which, given some parameters, provides all possible complete subgraphs.

OVRGROUP differs from REGROUP in three essential points :

- recurrent groups determined by OVRGROUP are, possibly but not necessarily, overlapping following a set of internal rules;
- OVRGROUP allows indicator species, *i.e.* species belonging to only one group, to be distinguished from ubiquitous or transition species;
- due to an intrinsic bias, REGROUP is irrelevant to handle situations with ubiquitous species, which OVRGROUP does well.

On the other hand, OVRGROUP differs from KDEND in several points :

- KDEND uses an ultrametric and does not provide an objective threshold value for the similarity;
- the linkage parameter, k, defining the overlap in KDEND, is arbitrarily fixed by the user (*);
- if no overlap occurs, KDEND becomes single linkage clustering.

This means that, contrary to KDEND, OVRGROUP provides a

(*) Practically, (k - 1) defines the maximum numbers of points in the overlap between the classes that belong to it. There is no overlap when k = 1.

solution which is unique and that the degree of overlap between groups is fixed through a set of internal rules.

Together with KDEND, OVRGROUP has to be classified among the clumping techniques (CORMACK, 1971; EVERITT, 1974), *i.e.* among cluster analysis techniques in which the classes or clumps can overlap. However, this paper does not present a general algorithm likely to solve very special problems as does KDEND. It merely proposes a defined, repeatable method likely to group together species which are frequent components of each other's biological environment and should be interpreted as an extension of FAGER's concept and algorithm.

ACKNOWLEDGEMENTS

My sincere thanks are due to Dr. Ph. LEBRUN under whom this work was achieved and to Drs. E.L. VENRICK (La Jolla, CA) and J.A. WALLWORK (London, UK) for reading and criticizing the manuscript. I am also grateful to Dr. VENRICK for providing the set of programs dealing with recurrent groups.

APPENDIX

The complete study of recurrent groups requires 4 to 5 programs :

AFFIN computes the treshold value T.
REGROUP and OVRGROUP form recurrent groups (exclusive or overlapping depending on the program).
CONNEX determines the relationships between groups previously defined.
STATION searches for releves where recurrent groups are present.

AFFIN, REGROUP, CONNEX and STATION are available from Dr. VENRICK, Scripps Institution of Oceanography, la Jolla, CA 92093. OVRGROUP and a new version of CONNEX (dealing with overlapping groups) available from the author.

REFERENCES

ANDRE H.M. (1985). Arthropod overlapping recurrent group in corticolous epiphytes : A case study. *(in prep.)*.

BROCKMANN-JEROSCH H. (1907). *Die Pflanzengesellschaften der Schweizeralpen. I. Die Flora des Puschlav und ihre Pflanzengesellschaften*. Engelman, Leipzig.

COLE A.J. and WISHART D. (1970). An improved algorithm for the Jardine-Sibson method of generating overlapping clusters. *Computer Journal* 13, 156-163.

CORMACK R.M. (1971). A review of classification. *Journal of the royal statistical society* 134, 321-367.

DALE M. (1977). Graph theoretical analysis of the phytosociological structure of plant communities : the theoretical basis. *Vegetatio* 34, 137-154.

DUVIGNEAUD P. (1946). La variabilité des associations végétales. *Bulletin de la Société royale de Botanique de Belgique* 78, 107-134.

ELLENBERG H. (1950). *Landwirtschaftliche Pflanzensoziologie. I. Unkrautgemeinschaften als Zeiger für Klima und Boden*. Verlag E. Ulmer, Stuttgart, 141 pp.

EVERITT B. (1974). *Cluster Analysis*. Heinemann, London, 122 pp.

FAGER E.W. (1957). Determination and analysis of recurrent groups. *Ecology* 38, 586-595.

FAGER E.W. and McGOWAN J.A. (1963). Zooplankton species groups in the North Pacific. *Science* 140, 453-460.

GAMS H. (1918). Prinzipienfragen der Vegetationsforschung. Ein Beitrag zur Begriffsklärung und Methodik der Biocoenologie. *Naturforschende Gesellschaft in Zurich* 63, 293-493.

GISIN H. (1946). Les groupements animaux dans leurs rapports avec le milieu. *Compte rendu de la Société de Physique et d'Histoire naturelle de Genève* 63, 45-46

GRADMANN R. (1909). Ueber Begriffsbildung in der Lehre von den Pflanzenformationen. *Botanisches Jahrbücher* 43, 91-103.

HUBÁLEK Z. (1982). Coefficients of association and similarity, based on binary (presence - absence) data : an evaluation. *Biological Review* 57, 669-689.

JANSON S. and VEGELIUS J. (1981). Measures of ecological associations. *Oecologia* 49, 371-376.

KLIR G.J. (1969). *An Approach to General Systems Theory.* Van Nostrand Reinhold Co , New York, 323 pp.

KRYLOV B.V. (1968). Species association in plankton. *Oceanology* 8, 243-251.

KÜHNELT W. (1969). *Ecologie générale*. Masson, Paris, 359 pp.

LEGENDRE P. and BEAUVAIS A. (1978). Niches et associations des lacs de la Rasissonie québécoise. *Naturaliste Canadien* 105, 137-158.

MARGALEF R. (1968). *Perspectives in Ecological Theory.* The University of Chicago Press, Chicago , 111 pp.

OCHIAI A. (1957). Zoogeographical studies on the soleoid fishes found in Japan and its neighbouring regions. *Bulletin of the Japanese Society of scientific Fisheries* 22, 526-530.

SHIMWELL D.W. (1971). *The Description and Classification of Vegetation*. Sidgwick and Jackson, London, 322 pp.

VENRICK E.L. (1971). Recurrent groups of diatom species in the North Pacific. *Ecology* 52, 614-625.

WHITTAKER R.H. (1962). Classification of natural communities. *Botanical Review* 28, 1-239.

WHISHART D. (1978). *CLUSTAN, user manual* (3rd ed.). Program Library Unit, University of Edinburgh, 175 pp.